Analysis of a Pilot Study for Amelioration of Itching in Liver Disease: When is a Failed Trial not a Failure?

Monnie McGee* & Nora V. Bergasa**

*Department of Statistical Science, Southern Methodist University
Dallas, TX 75275

**Division of Hepatology, SUNY Downstate Medical Center
Brooklyn, New York

Authors' footnote

Monnie McGee is Assistant Professor, Department of Statistical Science, Southern Methodist University, Dallas TX, 75275-0332 (mmcgee@smu.edu). Nora Bergasa is Professor of Medicine at the State University of New York at Downstate Medical Center, Kings County Hospital Center, Brooklyn, New York (Nora.Bergasa@downstate.edu). The authors acknowledge helpful comments from Richard Gunst, Lynne Stokes, an associate editor, and two anonymous reviewers while this manuscript was in preparation.

1. INTRODUCTION

Pilot studies and early phase clinical trials (Phase I and sometimes Phase II) are exploratory experiments conducted more to generate hypotheses than to confirm efficacy of a particular drug. The measure of "success" in such a trial is in the additional information gathered about the substance or disease under study, and whether that information is useful in generating new avenues of research. This paper describes, in detail not possible in medical journals, the application of advanced statistical methods to data generated by a randomized double-blind placebo-controlled pilot study of gabapentin for its efficacy in reducing scratching, secondary to itching, in patients with liver disease (Bergasa *et. al.*, to appear).

The data generated by the trial (and in most clinical trials) contain a variety of problems that are not typically discussed in statistics textbooks. Such issues include, but are not limited to, small numbers of subjects, missing data, and outliers. Missing values in a trial that is already small are of particular concern, since one does not want to discard any data and therefore decrease further the efficiency of any estimates. The presence of large outliers raises the question of whether to believe the data (*i.e.* the outliers are true values, and should be included in all analyses) or to delete or downweight them in order to fit a model. The small sample size implies that many of the asymptotic results dealing with the consistency of model estimates from methods such as multiple imputation are no longer applicable.

However messy the final measurements and the myriad of methodological problems they pose for the statistician, such results represent years of work for the principal investigators, and a significant time investment for the research subjects. Practicing statisticians face this quandary daily: how to glean all information possible from the data without sacrificing statistical principles, and to do so as quickly as possible. The story of this search is not often told in the resulting papers in subject-matter journals. Part of the purpose of this paper is to provide information about particular methods that might be helpful to a practicing statistician with a messy data set to analyze. The methods presented here are by no means the only tricks that can be used for dealing with messy data; they are given as an example of the effort required in analyzing a "simple" pilot study. The other, more important, purpose of the paper is to emphasize that the statistical results from a trial cannot always be interpreted by a statistician alone. Collaboration with an expert in the subject matter of the trial can lead to new theory to be tested.

2. THE GABAPENTIN TRIAL

The protocol called for sixteen subjects to be randomized to either gabapentin or a placebo. Before the treatments began, and prior to randomization, baseline data of scratching activity and perception of itch were obtained hourly over at least a 48-hour period in the hospital. After the initial (pretreatment) quantification, subjects were given their randomly assigned medication and asked to resume their normal daily routines. After four weeks on the study medication, the subjects returned to the hospital for a second 48-hour evaluation (called "on-treatment" in the sequel), which was conducted in the same way as the first.

Scratching activity, a behavior which results from the perception of itch, was measured by a monitoring system that consists of a piezo film sensor glued to a cast custom made to fit the middle finger on the dominant hand of the user (Talbot, *et. al.*, 1991). The main component of the system is a signal processor, which consists of a frequency counter incorporating a threshold detector and a bandpass filter to prevent extraneous counts from being registered. At the start of the recording session, subjects were asked to scratch over a defined distance on a piece of cardboard as if they were scratching their skin. The counts are added and recorded as hourly scratching activity (HSA). The result is a numerical value, which purports to be an objective measure of scratching behavior for each subject.

Large values indicate more scratching.

Any background noise recorded in the system was subtracted from the hourly scratching activity value. This subtraction resulted in some negative values, which were recorded as zeros. A value of zero means that the amount of scratching during that hour was below the background level of body movement, or that it was absent after movement. For any subject, the important thing about the measurement is the relative size of the pre-treatment value versus the on-treatment value. For these reasons, the zeroes were not treated as detection limits, instead, they were left in the data as real values.

Perception of itch was measured using a visual analog scale. The visual analog scale has been is used extensively in medical experiments as a way to measure outcomes such as pain and fatigue (Auburn, 2003; Hartmannsgruber, 2000; Wewers, 1990). The visual analog scale consists of a straight line, 10 millimeters in length. For the purposes of the gabapentin study, the left endpoint of the line represents “no itching” and the right endpoint “severe itching”. Subjects were asked to make a mark on the line corresponding to the severity of the itching they perceived to be experiencing at the time the instrument was administered. The distance from the left endpoint to the mark (in mm) is the visual analog score (VAS). For this study, VASs were measured every hour for at least 48 hours during baseline and post-treatment quantification periods.

3. EXPLORATORY DATA ANALYSIS

Statisticians recognize that real data are usually not well-behaved, and that proper preparation of the data is essential in obtaining a valid and reliable statistical analysis. In other words, it is necessary to perform an exploratory analysis in an attempt to obtain valid statistical inference from a small data set. Figure 1 shows baseline HSA values for patients in the gabapentin and placebo groups at seven time points during the study. Each plotting symbol (either a circle or a diamond, depending on the group) represents one patient. Initially, there are seven subjects in each group; data for one patient in the gabapentin group is not shown because the measuring device for that patient failed at the baseline quantification.

Figure 1 about here.

From Figure 1, we see that, at every time point, several subjects have very low HSA values. Some of these values are identical to zero. Second, for almost any parametric statistical analysis chosen for these data, stable model parameter estimates would be very difficult to obtain due to the presence of extremely large values. This is particularly true at hour 8, and is similarly pronounced for other hours that are not shown in this graph. An examination of patient behavior versus the times at which outliers occurred revealed that such large values were unexpected within the range of outputs from the piezo film sensor; therefore, all outliers were deleted and imputed using MNNHDI, described in the next section.

Finally, Figure 1 shows a high rate of missing data, as there are progressively fewer patients as the length of the hospital stay increases. This is not an artifact of the plotting procedure, as Table 1 shows. In this table, we see the percentage of values missing for each subject, for both the HSA and VAS measurements at both treatment times, previous to the deletion of outliers in the HSA measurements. Ideally, each subject should have had 48 pre- and on-treatment measurements for both variables. In reality, only two subjects completed a 48-hour recording session as originally designed; most subjects completed between 20 and 25 hours of recording. There are typically higher rates of missing values for VAS due to a provision in the study protocol which prevented researchers from disturbing the subjects while they were sleeping. Thus, approximately one-third of the VAS measurements are missing by design.

Table 1 about here

If this study were to be analyzed using a complete case analysis (CC), where only cases with complete data are used, the large number of missing values would seriously compromise the power of any analysis. A more promising alternative would be to replace all missing observations (as well as outliers) in a principled way.

4. IMPUTING MISSING DATA

There is a rich literature on imputation of missing data. Many articles dealing with imputation assume that data are missing completely at random (MCAR). Most of the time, the true mechanism is missing not at random (MNAR), also called non-ignorably missing (Carpenter and Kenward, 2005). In such cases, the desired method of dealing with the missing data involves modeling the missing data mechanism, which is typically unknown (Little and Rubin, 2002, Chapter 15) .

The missing observations from the gabapentin trial are generated by a mixture of mechanisms. For some values, we know that the mechanism generating missing values is independent of the observations themselves. For example, two of the subjects were missing entire pre- or on-treatment quantifications, because the machine recording HSA failed. The data missing for these two subjects would be considered MCAR. However, the same mechanism is not operating for all missing values. If we delete outliers, we will be doing so because they are too large. The mechanism generating those missing values will be MNAR by construction.

There is a growing literature on imputation of MNAR data for longitudinal studies, of which the pretest-posttest study is a special case. Reams and Van Deusen (1999) suggested using hot-deck imputation for estimating values missing by design in an annually-conducted forest inventory survey. The authors do not mention the effect of inherent correlation structure on the variance and bias of the imputed values. Liu and Gould (2002) evaluate last observation carried forward (LOCF), CC, multiple imputation (MI), and a mixed-effects model as replacement strategies for replacing missing data in longitudinal trials. They found that MI tended to perform the best for data MNAR, because of its flexibility to include surrogate variables, correlated with the missing values. However, the theoretical and empirical results from these articles rely on simulations and/or data sets where the number of subjects is many times that of number of subjects in the gabapentin study. In addition, these studies did not consider the situation where the variables measured at each time point in the study were time series themselves.

Since we cannot afford to throw out missing values, we need a sensible way to replace them. Missing observations for the gabapentin data were treated in two stages. First, the last 24 hours of the HSA measurements were deleted. Only two subjects had complete 48-hour records for both pre- and post-treatment quantifications in this study. The original purpose in collecting 48 hours of observations was to test for the presence of a 24-hour (*e.g.* circadian) rhythm in scratching activity, and imputation of the last 24 hours of observations would probably not yield usable estimates of any such rhythm. Without the last 24 hours of measurements, the most important part of the study can still be salvaged: to make a decision about the effectiveness of gabapentin.

There were four subjects with incomplete HSA data for the first twenty-four hours at the baseline assessment and five subjects with incomplete data for the first twenty-four hours at the on-treatment assessment (not accounting for missing data caused by outlier deletion). The five subjects missing on-treatment assessments include two subjects (Patients 8 and 9) who did not complete the study; therefore, they have no data at all for the on-treatment quantification. The equipment failed for a third subject (Patient 13); she is missing all on-treatment measurements, as well. After the last 24-hours of data were deleted, some VASs were missing for four subjects at the baseline quantification, and for ten subjects (including subjects 8 and 9) at the on-treatment quantification. Most of the missing values were the result of adherence to the study protocol stipulation that subjects not be disturbed while sleeping.

In the second stage, the outliers were determined by examining studentized residuals from a complete data (pre-imputation, last 24-hours deleted) analysis using the mixed-effects model given in Equation 3 (data not shown). The outliers were deleted and all missing observations were replaced by an observation from a matching subject (called a "donor"). This type of imputation is sometimes called "Nearest Neighbor" hot-deck imputation (NNHDI in the sequel) (Little and Rubin, 2002, p. 69).

More precisely, let $y_i = (y_{i1}, \ldots, y_{ik})$ be a $K \times 1$ complete–data vector of outcomes. Further, let $y_i = (y_{\text{obs},i}, y_{\text{obs},m})$ where $y_{\text{obs},i}$ is the observed part and $y_{\text{obs},m}$ is the missing part of y_i. Then

$$\hat{y}_{it} = y_{\ell t} + (\overline{y}_{\text{obs},i} - \overline{y}_{\text{obs},\ell}) \tag{1}$$

where $\overline{y}_{\text{obs},i}$ is the mean of the observed values for subject i. Subject ℓ is the donor.

It is important to choose a donor that is "close" to the subject whose observations are missing. "Close" is defined by a metric, (*e. g.* $d(i,j) = \max_k |x_{ik} - x_{jk}|$) where $x_i = (x_{i1}, \ldots, x_{iK})^T$ are the values of K appropriately scaled covariates for a unit i at which y_i is missing (Little and Rubin 2002, p. 69). For time series data, the distance metric is somewhat different. Suppose subject i is missing a value at time t. For this trial, the donor is defined as

$$d_j(t) = \min_j \sum_{t=1}^{T} |x_{it} - x_{jt}|, \tag{2}$$

for all $j = 1, \ldots, n-1$. Note that there are relatively few donors for the recipients in this study. Efforts were made to ensure that same donor was not used repeatedly. If one donor was chosen for two or more recipients, the next-nearest donor was substituted.

Donor subjects should be selected using another variable (the donor variable) besides the variable which is being imputed (recipient variable). The effect of correlation between donor variable and the recipient variable is addressed in the next section. For these data, the only other available variable is the VAS; therefore, nearest neighbors are determined by computing the distance (2) between the recipient and all other subjects (candidate donors) within the same treatment group using VAS measurements. Nearest neighbors were estimated using observed values; further possibilities for using measurements missing by design as donors are considered in the discussion.

Once a donor was selected, his or her HSA values were used to substitute for missing HSA observations in the recipient. In typical hot-deck imputation, the missing observations are replaced with donated observations only once, and the new data are used as the "real" data set. This provides no estimate of imputation error, nor does it reflect the variability between subjects. For example, even if two subjects have exactly the same VAS trajectory, it is quite likely that their HSA values will be different due to random variation. It is necessary to estimate the uncertainty associated with replacement of missing values.

Two modifications were made to NNHDI. First, a random perturbation was added to mimic the inherent variability in the data. The random perturbation is generated from a $\mathcal{N}(0,29)$ distribution. The variance of the additive noise is the variance of the middle 80% of the extant HSA observations calculated over all subjects. In addition, three sets of imputed values are obtained, with three different sets of donors. Three sets provide enough information to estimate the imputation uncertainty, while keeping the analysis simple.

5. SIMULATION EXPERIMENTS

In this section, two simulation experiments are described. The first experiment examines the effect of outlier presence on the correlation structure of an AR(1) process. The second examines effect size for a scenario where the HSA measurements from each subject are uncorrelated. For all simulations, 250 replications for $n = 15$ subjects and $t = 24$ time points per subject were computed, in order to mimic the gabapentin data. Fifty percent of the on-treatment values only were deleted to simulate the missing data.

The effect of the presence of outliers on the within subject correlation structure was evaluated by simulating AR(1) processes with two possible values of the coefficients: $\phi = 0.25$ or 0.5. These coefficients are reasonable based on an examination of the autocorrelation structure of the extant gabapentin data. Values that were k times greater than the mean of the process (where $k = 3$ or 6) were substituted into the series to represent outliers. Series with both one and two outliers were simulated. The results are given in Table 2.

Table 2 about here.

Even one outlier that is approximately three times larger than the mean of the series affects the researcher's ability to estimate correctly the order of the process and its coefficient. The fourth column in the table, labeled "% Correct Order" gives the percentage of times that the AIC criterion judged the AR(1) series to have order 1. Most of the time, the AIC picked 0 as the order of these series; sometimes the AIC selected orders as large as 5. In the fifth and sixth columns, it is evident that the presence of outliers results in poor estimates of the correlation coefficient. This simulation gives evidence that it is reasonable to delete outliers and replace them with imputed values.

We further examined the effect of the degree of correlation between the donor and recipient variables when replacing outliers and missing values. The pre- and on-treatment values were simulated as Gaussian white noise in order to examine the effect of between variable correlation without the confounding effect of within variable correlation. Donor variables were simulated such that the correlation coefficient between them and the recipient was either $\phi = 0$, $\phi = 0.25$, or $\phi = 0.75$. The results are evaluated in terms of the mean of the difference of the pre

HSA values and the post HSA values and the associated mean-squared error (Table 3).

Table 3 about here.

Even with 50% of values missing, replacement of those values with MNNHDI results in effect sizes that are quite close to the true effect sizes. When the donor variable is correlated with the recipient, the MSE improves, as would be expected. However, MNNHDI seems to be a reasonable replacement strategy even when the donor and recipient variables are uncorrelated.

6. MIXED–EFFECTS MODEL ANALYSIS

Now that we have evidence that the replacement strategy produces reasonable estimates of effect sizes, we turn to an analysis of the data using a mixed effects model. Using this model with the imputed data, we can determine whether gabapentin has an effect on HSA.

The gabapentin experiment, with hourly scratching activity measurements aggregated into average pre- and on-treatment measurements, can be seen as a split-plot design, where treatment (gabapentin or placebo) is the whole-plot factor, subjects are the whole plots, and the quantification time (baseline or post-treatment) are split-plot measurements. An equivalent analysis would be to consider this a repeated measures design, with baseline and post-treatment HSA scores being the repeated measures.

The model is given by

$$\mathbf{y_{ijk}} = \alpha_{\mathbf{i}} + \mathbf{b}_{j(i)} + \gamma_{\mathbf{k}} + (\alpha\gamma)_{ik} + \epsilon_{\mathbf{ijk}}, \tag{3}$$

where y_{ijk} is the response for the j^{th} subject in the i^{th} group at the k^{th} quantification. The fixed effect, α_i, $i = 1, 2$, represents the effect of treatment group; b_j, $j = 1, 2$ is a random effect for the j^{th} subject nested within the i^{th} group, with $b_{j(i)} \sim NID(0, \sigma_b^2)$; γ_k is a fixed effect of the k^{th} quantification, $k = 1, 2$; $(\alpha\gamma)_{ik}$ represents the fixed interaction effect between the i^{th} treatment and the k^{th} quantification, and $\epsilon_{ijk} \sim NID(0, \sigma^2 I)$.

The results in Table 4 are those of model 3 applied to the average of the three data sets modified via MNNHDI. The response variable is the logarithm of HSA. These results account for missing values and outliers in the data.

Table 4 about here.

A complete case analysis using model 3 resulted in highly significant effects for the group effect, the quantification effect, and their interaction. In addition to the bias incurred from use of complete cases only, outliers are also present in these data, and may account for differences in the effects. In Table 4, the group effect and the interaction of group and quantification are still significant, but the quantification effect is no longer significant. These results indicate that there are differences in the scratching activity between gabapentin and placebo groups, but those differences cannot be attributed to the time at which the HSA measurements were taken.

Nonresponse uncertainty may also account for some of the differences in the CC and imputed analyses. Little and Rubin (2002, pages 86-87) give a method for fraction of information about a parameter θ due to nonresponse (denoted γ). The larger the fraction, the more influence imputation has over the parameter estimates. It is applied here in order to obtain an idea of how much of the variability in the model can be attributed to the replacement of the missing values.

Let $\hat{\theta}_d$ and W_d, $d = 1, \ldots, D$, be D complete-data estimates and their associated variances for θ. $\hat{\gamma}_D = (1 + 1/D)B_D/T_D$ is an estimate of the fraction of information about θ due to nonresponse, where W_D is the within-imputation variance, B_D is the between-imputation variance, and T_D is the total variability across imputations. In the case of the gabapentin analysis, $D = 3$, and $\hat{\gamma}_D$ was less than one percent for the estimates of LME coefficients for group, quantification, and the interaction term. For the random effect of subject within group, approximately 52% of the information is due to nonresponse. This implies that the inferences made from the imputed data for the fixed effects can be trusted. As for the random effect, more data and further analyses are needed before its importance can be ascertained.

7. DISCUSSION

Statistical practice is never as neat as textbooks sometimes imply. The challenge for a statistician is not only to analyze the data in a reasonable manner, but to obtain reasonable data to analyze. A statistician rarely has the luxury of declaring a trial a failure once the treatments have been administered and the measurements have been collected. It is a statistician's job (and joy?) to extract as much information as possible from weak data, while remaining true to statistical theory and practice. For these data, missing observations were replaced using a modified nearest-neighbor hot deck imputation (MNNHDI). Outliers were also deleted, and subsequently replaced in the same manner. Simulation studies suggest that a mixed–effects analysis produces reliable parameter estimates.

Simulation studies also showed that the presence of outliers, and, by extension, the replacement of those outliers with imputed data, affects the correlation structure of measurements gathered over time. For this study, this effect was irrelevant, as the purpose of imputation was to obtain more reasonable estimates of pre- and on-treatment means in order to analyze the data with a mixed effects model. The use of MNNHDI in situations in which it is important to maintain the correlation structure in the presence of missing data remains to be explored. Pfeffermann and Nathan (2002) discuss some imputation methods for time series, and develop a new method that applies in the case where data are missing in waves, meaning that several values are missing in a row. Different patterns of missingness were not considered for this study; however, this is an important direction for future research, particularly in the case where the number of subjects is small.

Another future direction for research would be to consider the impact of using CC measurements of a donor variable (e.g. VAS) for determining recipients. It is conceivable to use an iterative process, where the primary variable of interest (e.g. HSA) is first used to impute values for a secondary variable; then the imputed values of the secondary variable are used to obtain donors for the primary variable. The process could be repeated until a measure of imputation variance was below a certain threshold. Such "double imputation" (using imputed values to obtain more imputed values) has been used with some success in microarray analysis (Kim, *et. al.*, 2004).

Some would say that the gabapentin trial was a failed trial, because the evidence for the drug's effectiveness was inconclusive. There are several reasons that this might be true. First, any efficacy of gabapentin may have been lost due to measurement error in the device used. This could certainly be the case given the large variability in the scratching measurements. However, it is not likely, as the deletion and imputation of outliers reduced the variability in the outcome measurements so that reasonable parameter estimates could be obtained. In addition, the instrument used to measure scratching activity has been used successfully in a wide variety of trials (Bergasa, *et. al.*, 1992, 1995, 1999). The failure of the instrument twice during this trial was an anomaly, and could be remedied by an examination of the equipment between subjects in future trials.

Second, the statistical imputation method may be too biased toward the null so as to damp out effectiveness. However, it should be noted that MNNHDI attempts to use donor variables from the same treatment group as the recipient, in order to minimize such bias. Furthermore, the simulation studies described in the previous section give evidence to the contrary.

Finally, it may be that gabapentin is indeed ineffective in ameliorating itch. This is likely the truth, but for more complicated reasons than the absence of a satiating effect of the drug. The itch from liver disease is believed to result, at least in part, from increased neurotransmission by the endogenous opioid system in the brain. Dopamine is a chemical released in the brain that reduces the brain's ability to register unpleasant stimuli. It is possible that the placebo group may have had high endogenous dopamine release in anticipation of receiving gabapentin. The subjects in the gabapentin group may have had the same expectation, but the effect of the dopamine release was muted by the drug. As a result, the patients on active treatment tended to respond as placebo patients.

This study is a case in point of how a “negative” outcome of a trial can open new possibilities for cross-displinary research. A statistician is not likely to know about the effects of gabapentin on endogenous dopamine release, but consultation with an expert scientist resulted in a new line of inquiry, into the mechanisms responsible for the placebo effect within the human brain. In this sense, the failure of gabapentin to reduce the sensation of itching in these subjects was not a failure of the trial at all.

References

Aubrun, F., Langeron, O., Quesnel, C., Coriat, P., and Riou, B. (2003), “Relationships between Measurement of Pain Using Visual Analog Score and Morphine Requirements during Postoperative Intravenous Morphine Titration,” *Anesthesiology*, 98, 1415-1421.

Bergasa, N.V., Alling, D.W., Talbot, T.L., Swain, M.G., Yurdaydin, C., Schmitt, J.M., Walker, E.C., and Jones, E.A. (1995), “Naloxone ameliorates the pruritus of cholestasis: results of a double-blind randomized placebo-controlled trial,” *Annals of Internal Medicine*, 123, 161-167.

Bergasa, N.V., Alling, D.W., Talbot, T.L., Wells, M., Jones, E.A. (1999), “Oral nalmefene therapy reduces scratching activity due to the pruritus of cholestasis: a controlled study," *Journal of the American Academy of Dermatology*, 41, 431-434

Bergasa, N.V., Link, M.J., Keogh, M., Yaroslavsky, G., Rosenthal, R.N., McGee, M (2001), "Pilot study of bright-light therapy reflected toward the eyes for the pruritus of chronic liver disease," *American Journal of Gastroenterology*, 96, 1563-1570,

Bergasa, N.V., McGee, M., Ginsburg, I., and Engler, D., "Gabapentin Treatment for the Pruritis of Cholestasis: Results of a Double-Blind Placebo-Controlled Trial," *Hepatology* (to appear).

Bergasa, N.V., Talbot, T.L., Alling, D.W., Schmitt, J.M., Walker, E.C., Baker, B.L., Korenman, J.C., Park, Y., Hoofnagle, J.H., and Jones, E.A. (1992), "A controlled trial of naloxone infusions for the pruritus of chronic cholestasis," *Gastroenterology* , 102, 544-549.

Carpenter, J. and Kenward, M. (2005), Economic and Social Research Council Missing Data Website. http:www.missingdata.org.uk. Date of Access: May 17, 2005.

Hartmannsgruber, M.W.B. and Silverman, D.G. (2000), "Applying Parametric Tests to Visual Analog Scores", *Anesthesia & Analgesia*, 91, 248 - 249.

Kim, K.Y., Kim, B.J., and Yi, G.S. (2004), "Reuse of imputed data in microarray analysis increases imputation efficiency", *BMC Bioinformatics*, 5:160.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with*

Missing Data, 2nd ed.. New York: Wiley Interscience.

Liu, G. and Gould, A. L. (2002), “Comparison of alternative strategies for analysis of longitudinal trials with dropouts”, *Journal of Biopharmaceutical Statistics*, 12, 207-226.

Pfeffermann, D. and Nathan, G. (2002), “Imputation for Wave Nonresponse: Existing Methods and a Time Series Approach”, in *Survey Nonresponse* (Groves, R. M., Dilman, D. A., Eltinge, J. L., and Little, R. J. A., eds.). New York: Wiley, 417-430.

Reams, Gregory A. and Van Deusen, Paul C. (1999), “The Southern Annual Forest Inventory System”, *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 346-360.

Talbot, T.L., Schmitt, J.M., Bergasa, N.V., Jones, E.A., and Walker, E.C. (1991), “Application of Piezo Film Technology for the Quantitative Assessment of Pruritis”, *Biomedical Instrumentation and Technology*, 25, 400-403.

Wewers, M.E. and Lowe, N.K. (1990) “A Critical Review of Visual Analog Scales in the Measurement of Clinical Phenomena”. *Research in Nursing and Health*, 13, 227-236.

		HSA		VAS	
Group	Subject	Baseline	Final	Baseline	Final
Gabapentin	1	52	46	60	52
	2	52	44	60	67
	6	40	0	48	23
	8	44	100	50	100
	10	0	0	25	33
	12	100	46	17	54
	14	2	2	21	25
	16	21	79	100	88
Mean for Gabapentin Group		39	40	48	55
Placebo	3	0	0	31	40
	4	4	8	38	42
	7	0	44	25	58
	9	46	100	63	100
	11	50	19	19	52
	13	94	100	27	67
	15	10	52	29	63
Mean for Placebo Group		29	46	33	60
Overall Mean		34	43	41	58

Table 1: Percent of observations missing for each subject at each treatment period for both HSA and VAS variables, prior to deletion of the last 24-hours of the quantification period. Approximately 30 percent of the VAS measurements are missing by design, and are included in the percents given.

True ϕ	k	No. of outliers	% Correct Order	Estimated ϕ	MSE
0.25	3	1	28	0.171	0.044
0.25	6	1	10	0.037	0.136
0.5	3	1	50	0.045	0.316
0.5	6	1	18	0.046	0.296
0.25	3	2	20	0.047	0.143
0.25	6	2	16	0.055	0.146
0.5	3	2	51	0.054	0.305
0.5	6	2	46	0.071	0.267

Table 2: Effect of deletion of outliers from AR(1) series on the estimated order and the estimated coefficient. Outliers were defined as values that were k times the mean of the series. The table gives the percentage of times the AIC criterion gave the correct order, the estimated value of ϕ, and the MSE for the estimate.

$n = 15$ with eight missing values			
True Effect	Correlation	Imputed Effect	MSE
0	None	0.004	0.0989
	0.25	0.0008	0.0911
	0.75	0.0005	0.0756
3	None	2.996	0.0991
	0.25	3.001	0.0911
	0.75	3.003	0.0755

Table 3: Effect sizes where white noise data is replaced by MNNHDI. True Effect is the intended difference between the means of the pre- and on-treatment data, Correlation is the correlation between the donor and recipient variables, Imputed Effect is the effect size once on-treatment values were replaced by MNNHDI, and the MSE gives the mean squared error of the imputed effect.

Effect	DF	Mean Square	F Value	Pr > F
Group	1	1.74	8.29	0.0129
Subject (Group)	13	0.309	1.47	0.2480
Quant	1	0.469	2.23	0.1588
Group × Quant	1	1.32	6.31	0.0260

Table 4: Results for Average of 3 Imputations of NNHDI with additive $\mathcal{N}(0, 29)$ noise.

Figure Caption

Strip chart of HSA values for patients in Gabapentin (dark gray diamonds) and Placebo (black circles) groups at hours 1, 8, 16, 24, 32, 40, and 48 of baseline quantification.

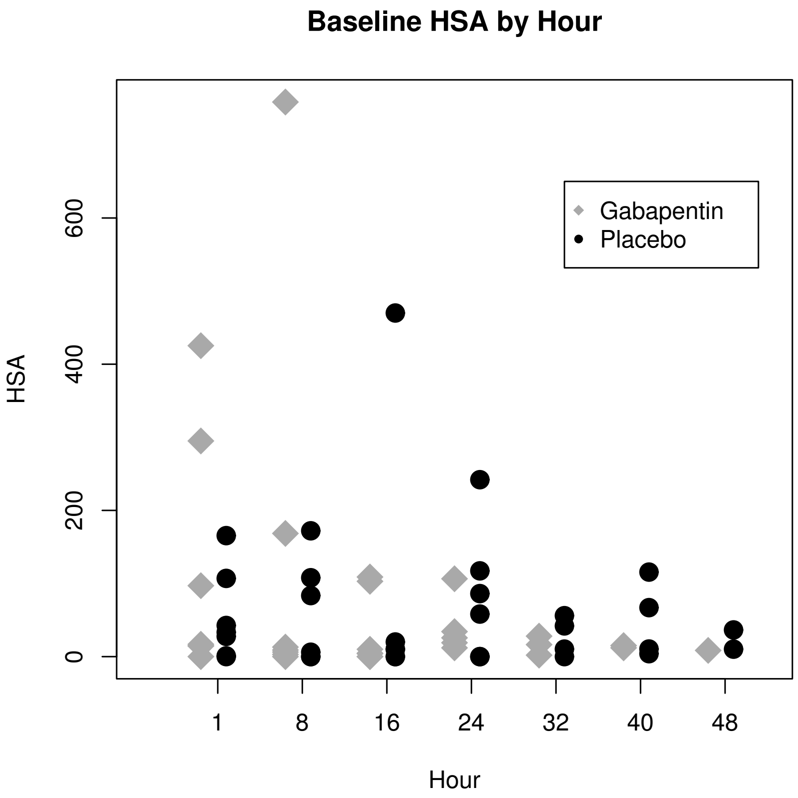

Figure 1: